Stat 414 - Day 11
Random Slopes cont.

Last Time:

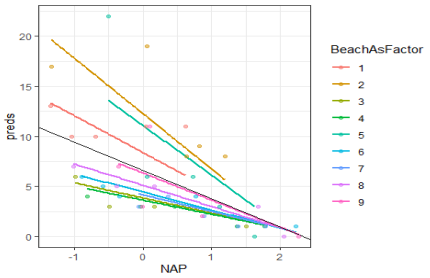
- Interaction terms change slopes
  - Interpret "main effects" by "zeroing out" the interaction term.
  - Otherwise, coefficient of $x_1 = \hat{\beta}_1 + \hat{\beta}_3 x_2$
  - Use product terms, centering quantitative variables can reduce multicollinearity
- We fit a "random coefficient model" $y_{ij} = \beta_{0j} + \beta_{1j} x_{1ij} + \epsilon_{ij}$ where $\beta_{0j} = \beta_{00} + u_{0j}$ with $u_{0j} \sim N(0, \tau_0^2)$ and $\beta_{1j} = \beta_{10} + u_{1j}$ with $u_{1j} \sim N(0, \tau_1^2)$
  - So β_{1j} are normally distributed around β_{10}
  - This adds a variance component for the slopes (τ_1^2) as well as a covariance between the random slopes and random intercepts (τ_{01})

Example 1: Continuing with our beach data Adding random slopes plays the role of an interaction between the Level 1 variable and the Level 2 "grouping variable." In this case, we tried a random slopes model, and the variation in the slopes (τ_1^2) was statistically significant.

```
#library(lme4)
rikzdata <- read.table("http://www.rossmanchance.com/stat414/data/RIKZ.txt", header
=TRUE)
model0 = lmer(Richness ~ 1 + (1 | rikzdata$Beach), data = rikzdata)
model1 = lmer(Richness ~ NAP + (1 | Beach), data  = rikzdata)
model2 = lmer(Richness ~ NAP + (1 + NAP | Beach), data = rikzdata, REML = FALSE)
```

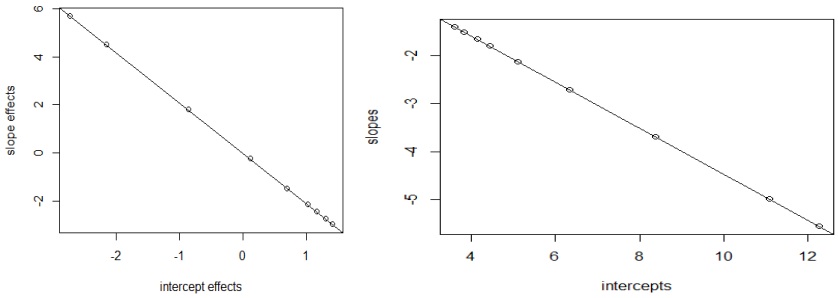
```
summary(model2)
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  Beach    (Intercept) 10.949   3.309
##           NAP          2.502   1.582    -1.00
##  Residual              7.174   2.678
## Number of obs: 45, groups:  Beach, 9
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   6.5818     1.1883   5.539
## NAP          -2.8293     0.6849  -4.131
```

```
ranef(model2)
## $Beach
##   (Intercept)        NAP
## 1   1.7986325 -0.8597860
## 2   5.6926249 -2.7212003
## 3  -2.7426699  1.3110567
## 4  -2.9682494  1.4188887
## 5   4.5044936 -2.1532473
## 6  -2.1372277  1.0216420
## 7  -2.4398536  1.1663039
## 8  -1.4646228  0.7001220
## 9  -0.2431276  0.1162204
```



(a) Between what two values do we expect 95% of the slopes to fall?

But we notice this gives us a second parameter as well. What does it mean for the slopes and intercepts to be correlated?

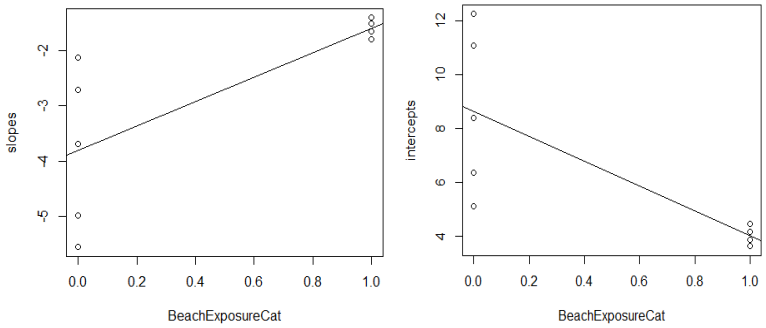


(b) Interpret the strong, negative correlation ($corr(u_{0j}, u_{1j}) = \hat{\tau}_{01}$ = -0.99) between the slopes and intercepts in context: beaches with larger intercepts (meaning what?) tend to have what kinds of slopes (meaning what?).

Recall we turned Exposure into a binary variable for high vs. low exposure (an index composed of the following elements: wave action, length of the surf zone, slope, grain size, and the depth of the anaerobic layer).

(c) Prediction: Now consider adding this Level 2 variable to the model. What do you expect to change in the model?

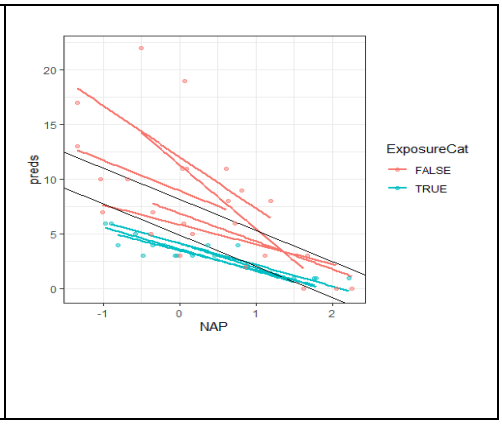
Explore whether exposure is related to the intercepts and/or the slopes.



(d) Is Exposure "positively" or "negatively" related to the intercepts? How about the slopes? What are the implications to the model?

Fit the model including Exposure

```
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  Beach    (Intercept) 5.371    2.318
##           NAP         2.681    1.637    -0.84
##  Residual             6.756    2.599
## Number of obs: 45, groups:  Beach, 9
##
## Fixed effects:
##                 Estimate Std. Error t value
## (Intercept)       8.1923     1.0567   7.753
## NAP              -2.8516     0.6931  -4.114
## ExposureCatTRUE  -3.3238     1.2795  -2.598
```



(e) Interpret the slope coefficient of Exposure in this model.

(f) Is the Exposure variable significant? Do you see a substantial improvement in the fit of the model? How do the variance components change/what has been the main impact?

(g) Why didn't including the Exposure variable explain much variation in the slopes?

(h) To expand our model to allow for Exposure to explain variation in slopes, write the Level 1 and Level 2 equations, including Exposure in both Level 2 equations.

(i) Now make the composite equation, what happens? (Hint: What is the expression for the intercepts and what is the expression for the slopes?)

```
model4 = lmer(Richness ~ NAP*ExposureCat + (1 + NAP | Beach), data  = rikzdata,
REML=FALSE)  #with interaction

## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  Beach    (Intercept) 3.832    1.957
##           NAP         1.002    1.001    -1.00
##  Residual             7.161    2.676
## Number of obs: 45, groups:  Beach, 9
##
## Fixed effects:
##                     Estimate Std. Error t value
## (Intercept)           8.9590     1.0474   8.553
## NAP                  -3.8812     0.7228  -5.370
## ExposureCatTRUE      -5.3824     1.5864  -3.393
## NAP:ExposureCatTRUE   2.4460     1.0991   2.225
```

```
## ============================================================================
##                             no exposure       no interaction    interaction
## ----------------------------------------------------------------------------
## (Intercept)                   6.582 (1.188)     8.192 (1.057)     8.959 (1.047)
## NAP                          -2.829 (0.685)    -2.852 (0.693)    -3.881 (0.723)
## ExposureCatTRUE                                -3.324 (1.280)    -5.382 (1.586)
## NAP:ExposureCatTRUE                                               2.446 (1.099)
## ----------------------------------------------------------------------------
## AIC                          246.656           245.335           243.221
## BIC                          257.496           257.982           257.674
## Log Likelihood              -117.328          -115.668          -113.611
## Num. obs.                     45                45                45
## Num. groups: Beach             9                 9                 9
## Var: Beach (Intercept)        10.949             5.371             3.832
## Var: Beach NAP                 2.502             2.681             1.002
## Cov: Beach (Intercept) NAP    -5.234            -3.203            -1.959
## Var: Residual                  7.174             6.756             7.161
## ============================================================================
anova(model2, model3, model4)
## Data: rikzdata
## Models:
## model2: Richness ~ NAP + (1 + NAP | Beach)
## model3: Richness ~ NAP + ExposureCat + (1 + NAP | Beach)
## model4: Richness ~ NAP * ExposureCat + (1 + NAP | Beach)
##        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model2    6 246.66 257.50 -117.33   234.66
## model3    7 245.34 257.98 -115.67   231.34 3.3207  1    0.06841 .
## model4    8 243.22 257.67 -113.61   227.22 4.1143  1    0.04252 *
```

(j) How many parameters did we add to the model? What is the estimate for that parameter? Is it statistically significant? How are you deciding?

Notes:

- Keep in mind that the variance of the intercepts is at "x = 0"
- With a level 2 variable, we are thinking of the Level 1 intercepts and slopes as "outcomes" and then running a regression model to explain that variation
- One recommended approach for model selection is to start with all potential fixed effects (including interactions), and decide on the random effects (e.g., slopes and/or intercepts). Then use that model to pare down the fixed effects.
- "In cases where the explanation of the random effects works extremely well, one may end up with models with no random effects at level two… random intercepts, slope have zero variance.. Omitted.. The resulting model may be analyzed just as well with OLS regression analysis… within group dependence has been fully explained by the available explanatory variables/interactions (no more dependence in the residuals)."

Computer Problem 11 - due Friday 7am

(k) How much variability in the slopes have we explained by including the interaction term?

(l) How did the residual standard error (or within group variance estimate) change? What does that tell you?

(m) Does Model 4 appear to be a better fitting model? How are you deciding? (Hint: What measures of "model fit" do we have that we haven't done much with recently?)

(n) How do you interpret the coefficient of the interaction term?

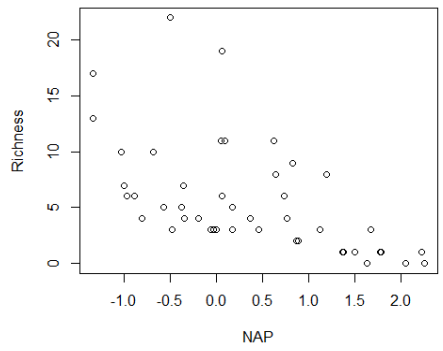
(o) In particular, give rough estimates of the overall intercept and the overall slope for low exposure beaches, and rough estimates for the overall intercept and the overall slope for high exposure beaches.

(p) Summarize what you learn from this model and how it illustrates the above observations. In particular, what are the black lines? What are the peach and blue points?

Consider the following two models

(q) Write a few sentences explaining the differences in these two models, what are they assuming/how they are modelling the data. Which model would you recommend and why?

Example 1 cont: Here is the original scatterplot:



(a) Do you consider the homoscedasticity assumption met?

The nlme package allows us to see that variance-covariance matrix for each beach. Here is that matrix for the five observations in Beach 1, and then the correlation matrix.

```
vcm = getVarCov(model1, type = "marginal", individual = "1"); vcm
## Marginal variance covariance matrix
##         1       2       3       4       5
## 1 18.0300  8.6675  8.6675  8.6675  8.6675
## 2  8.6675 18.0300  8.6675  8.6675  8.6675
## 3  8.6675  8.6675 18.0300  8.6675  8.6675
## 4  8.6675  8.6675  8.6675 18.0300  8.6675
## 5  8.6675  8.6675  8.6675  8.6675 18.0300
##   Standard Deviations: 4.2461 4.2461 4.2461 4.2461 4.2461
cov2cor(vcm[[1]])
##           1         2         3         4         5
## 1 1.0000000 0.4807353 0.4807353 0.4807353 0.4807353
## 2 0.4807353 1.0000000 0.4807353 0.4807353 0.4807353
## 3 0.4807353 0.4807353 1.0000000 0.4807353 0.4807353
## 4 0.4807353 0.4807353 0.4807353 1.0000000 0.4807353
## 5 0.4807353 0.4807353 0.4807353 0.4807353 1.0000000
```

(b) What are the values along the diagonal of the vcm matrix? What are the off-diagonal values?

(c) What are the off-diagonal values after running cov2cor? How do we convert?

Now let's look at the random coefficients model (with lme):

(d) Explain how this model allows us to account for the heterogeneity we mentioned in (a).

Looking at the variance covariance matrix:

```
vcm2 = getVarCov(model2, type = "marginal"); vcm2
## Beach 1
## Marginal variance covariance matrix
##         1      2      3       4      5
## 1 19.3650 18.431 20.200  8.6936 16.356
## 2 18.4310 35.545 30.962 13.2500 25.046
## 3 20.2000 30.962 41.254 14.5150 27.458
## 4  8.6936 13.250 14.515 13.5920 11.766
## 5 16.3560 25.046 27.458 11.7660 29.522
##   Standard Deviations: 4.4005 5.962 6.4229 3.6867 5.4334
```

(e) According to the fitted model, is the variance constant? Which observations in Beach 1 have larger variance?

Examine the data for the 5 observations for beach 1:

```
head(rikzdata, 5)
##   Sample Richness Exposure    NAP Beach ExposureCat
## 1      1       11       10  0.045     1       FALSE
## 2      2       10       10 -1.036     1       FALSE
## 3      3       13       10 -1.336     1       FALSE
## 4      4       11       10  0.616     1       FALSE
## 5      5       10       10 -0.684     1       FALSE
```

(f) What is true about the NAP values for the observations with higher predicted variance? The smallest predicted variance? In other words, the variance in the predicted Richness values (increases/decreases) with NAP?

```
cov2cor(vcm2[[1]])
##           1         2         3         4         5
## 1 1.0000000 0.7025218 0.7146676 0.5358621 0.6840891
## 2 0.7025218 1.0000000 0.8085487 0.6028179 0.7731744
## 3 0.7146676 0.8085487 1.0000000 0.6129561 0.7867876
## 4 0.5358621 0.6028179 0.6129561 1.0000000 0.5873949
## 5 0.6840891 0.7731744 0.7867876 0.5873949 1.0000000
```

(g) According to the fitted model, is the correlation between two observations within beach 1 the same for any two observations, or does it vary depending on which two observations you are pairing? Identify two observations in beach 1 that are more highly correlated, and two observations in beach 1 that are less correlated. (Do you see a pattern in their NAP values?)

The point is that a random slopes model also allows us to model heterogeneity in the data (y_{ij}) and that the amount of correlation between two observations depends on the corresponding x_{ij} values.

On HW 6, you will show the variance is a quadratic function in NAP $\tau_0^2 + x_{ij}^2\tau_1^2 + 2x_{ij}\tau_{01} + \sigma^2$

(h) so is minimized at $x_{ij} =$

(i) What does τ_{01} *represent? Suggest 3 ways to find the estimate of this model from the output you have seen.*

(j) Find the value of NAP that minimizes $Var(y_ij)$ *for our fitted model. Is this a value in the range of our data?? (Does your answer agree with the graph of the model?)*

The idea is when the correlation between the intercepts and slopes is negative, the lines are "fanning in" and variability is smaller for larger x values. If the correlation between the slopes and intercepts is positive, then the lines will "fan out" and variability in y is increasing for larger x values. But also watch for the point where they switch from fanning in to fanning out… If the correlation is close to zero, then there is no fanning, and you will have a scatter of positive and negative lines.

You will shown in HW 6, that the covariance between two observations also depends on the x values: $Cov(y_{ij}, y_{kj}) = \tau_0^2 + (x_{ij} + x_{kj})\tau_{01} + +x_{ij}x_{kj}\tau_1^2$

(k) What happens to the covariance between two observations when NAP = 0 (for both observations)? What about the correlation?

Notes:

- Bottom line: the variance and covariance in our data (y_{ij}) values now depend on the x_{ij} values, but τ_0^2 represents the variation in the intercepts when $x = 0$ and $(\tau_0^2)/(\tau_0^2 + \sigma^2)$ is the correlation of two measurements on the same beach with $x = 0$.
- But in general now have "fanning lines" and it may not make sense to calculate ICC. Or do so conditional on a particular value of x. In general, be more detailed when talked about "variability explained."